FUNDAMENTAL PROPERTIES OF PROCESS DISTANCES

JULIO BACKHOFF VERAGUAS, MATHIAS BEIGLBÖCK, MANU EDER, AND ALOIS PICHLER

Abstract. Information is an inherent component of stochastic processes and to measure the distance between different stochastic processes it is often not sufficient to consider the distance of their laws. Instead, the information which accumulates over time and which is mathematically encoded by filtrations, has to be accounted for as well. The *nested distance/ bicausal Wasserstein distance* addresses this challenge. It is of emerging importance in stochastic programming and other disciplines.

In this article we establish a number of fundamental properties of the nested distance. In particular we prove that the nested distance of processes generates a Polish topology but is itself not a complete metric. We identify its completion to be the set of *nested distributions*, a form of generalized stochastic processes, recently introduce by Pflug.

Moreover we find that — choosing an appropriate underlying metric — the nested distance induces Hellwig's *information topology* studied in the economic literature and in particular our results lead to new insights also in this context.

Keywords: Optimal Transport, nested distance, causal Wasserstein distance, information topology, Knothe-Rosenblatt rearrangement.

AMS subject classifications. Primary 60G42, 60G44; Secondary 91G20.

1. Introduction

In this paper, we consider several distances between stochastic processes and investigate their fundamental metric and topological properties. The distances we discuss are based on transport theory; we refer to Villani's monographs [24, 25] or the lecture notes by Ambrosio and Gigli [2] for background. Classical transport distances (cf. [22, 23]) do not respect the information structure inherent to a multivariate distribution when this is seen as a stochastic process. It is therefore desirable to find natural extensions of these distances that do take information/ filtrations into account. To achieve this, one has to adjust the definition of transport distances and include constraints involving the filtration to incorporate information at specified times. Heuristically this means that the computation of the distance is done over transport plans/couplings that only *move mass* respecting the *causal* structure inherent to filtrations.

These ideas lead to the *nested distance* introduced by Pflug in [18], and its systematic investigation was continued in [19, 20, 17]. The nested distance has already turned out to be a crucial tool for applications in the field of multistage stochastic optimization, where problems can be computationally extremely challenging, and in many situations they simply cannot be managed in reasonable time. Based on the nested distance, approximation with tractable simplifications becomes feasible, and sharp bounds for the approximation error can be found. Independently, a systematic treatment and use of causality as an interesting property of abstract transport plans and their associated optimal transport problems was initiated by Lassalle in [15]; in particular he introduces a nested distance under the name *(bi)causal Wasserstein distance* and provides intriguing connections with classical geometric/ functional analytic inequalities, as well as stochastic analysis. In [5, 1] it is argued that the classical Knothe-Rosenblatt rearrangement (cf. [25, Introduction]), also known as quantile transform in statistics, is a causal analogue to the celebrated Brenier-mapping in optimal transport and show that causality is naturally linked to subject of enlargement of filtrations in stochastic analysis; these articles are in the wider tradition of constrained transport problems and in particular related to martingale optimal transport (cf. [8, 12, 11, 7, 16] among many others).

While the nested distances has already received some attention, a number of basic and fundamental questions on the corresponding topological/ metric structure have remained open. The main goal of the present article is to fill this gap. Although the nested distance is inspired by the Wasserstein distance, it turns out that there are substantial differences between these concepts. In Section 3 we show that

The second and third authors acknowledge support by the Austrian Science Fund (FWF) under grant Y782-N25.

convergence in nested distance cannot be verified by testing against a class of usual functions. Furthermore we observe that nested distance is an incomplete metric. Our first main result, Theorem 4.4, identifies the completion of this topology explicitly. This completion turns out to be the space of *nested distributions* (cf. [18] and Definition 4.2 below), which are in a sense generalized stochastic processes, equipped with a classical Wasserstein distance. We thus connect two hitherto unrelated mathematical objects in an unexpected way. Interestingly we find that while the nested distance is not complete, we will establish in Section 5 that the induced topology is Polish, i.e., separable and completely metrizable (cf. Theorem 5.6 and comments thereafter). This is the second main result of the article. As a consequence of these considerations we can moreover find a complete metric compatible with the nested distance.

Most naturally, the nested distance is defined as a variant of a transport distance where the cost function is based on the usual Euclidean distance on $\mathbb{R}^N$, but we want to argue that also the nested distance defined from a bounded metric on $\mathbb{R}^N$ can be of interest. In the classical setup, a bounded metric on $\mathbb{R}^N$ induces a Wasserstein distance corresponding to the weak topology on probability measures. Likewise a nested distance based on a bounded metric induces an information-compatible topology which might be seen as a *weak nested topology.* We will establish that this weak nested topology coincides with the *information topology* introduced by Hellwig [13] (cf. also [6] and the references therein) for applications in the field of mathematical economics, more specifically, sequential decision making and equilibria. Our results in Sections 4 and 5 seem to be novel in the setup of [13], and potentially applicable in mathematical economics. We stress that the concept of nested distance (in contrast to weak nested topology) is not present, nor is it related, to the works [13, 6] to the best of our knowledge.

As we already mentioned, the Knothe-Rosenblatt rearrangement is the causal analogue of Brenier's map. This rearrangement appears everywhere in mathematics, from statistics to the theory of geometric inequalities. For this reason we will also compare the nested distance to a new distance defined in terms of (i.e. induced by) this rearrangement, which a priori seems easier to compute. In dimension one it happens that both distances coincide. Our finding here is that, in higher dimensions, this new distance is strictly stronger than the nested distance. This leads us to conjecture that in multiple dimensions there is no privileged transport/rearrangement that may induce a *simpler* metric topologically equivalent to the nested distance one.

Outline. We introduce the notation used throughout the paper and describe the mathematical setup in Section 2. In Section 3 we discuss some elementary properties of the nested distance. Section 4 is concerned with the with completeness-properties of this distance. Then in Section 5 we introduce the weak nested topology, compare it to Hellwig's information topology, and establish its Polish character. In Section 6 we introduce the Knothe–Rosenblatt distance and provide a comparison with the nested distance. Finally, we give a brief summary of our results in Section 7.

2. Notation and mathematical setup

The pushforward of a measure γ by a map M is $M_*\gamma := \gamma \circ M^{-1}$. For a product of sets $\mathcal{X} \times \mathcal{Y}$ we denote by p^1 (p^2, resp.) the projection onto the first (second, resp.) coordinate. We denote by γ^x, γ^y the regular kernels of a measure γ on $\mathcal{X} \times \mathcal{Y}$ w.r.t. its first and second coordinate, respectively, obtained by disintegration (cf. [3]) so that $\gamma(A \times B) = \int_A \gamma^{x_1}(B)\, \gamma^1(\mathrm{d}x_1)$ with $\gamma^1(A) := p^1_*\gamma(A) = \gamma(A \times \mathcal{Y})$. The notation extends analogously to products of more than two spaces. We convene that for a probability measure η on $\mathbb{R}^N$, $\eta^{x_1,\dots,x_t}$ denotes the one-dimensional measure on x_{t+1} obtained by disintegration of η w.r.t. $(x_1, \dots, x_t)$. Also, a statement like "for η-a.e. $x_1, \dots, x_t$" is meant to denote "almost-everywhere" with respect to the projection of η onto the coordinates $(x_1, \dots, x_t)$. On $\mathbb{R}^N \times \mathbb{R}^N$ we denote by $(x_1, \dots, x_N)$ the first half and by $(y_1, \dots, y_N)$ the second half of the coordinates. Similarly, we use the convention that for a probability measure γ on $\mathbb{R}^N \times \mathbb{R}^N$, $\gamma^{x_1,\dots,x_t,y_1,\dots,y_t}$ denotes the two-dimensional measure on (x_{t+1}, y_{t+1}) given by regular disintegration of γ w.r.t. $(x_1, \dots, x_t, y_1, \dots, y_t)$, so a statement like "for γ-a.e. $x_1, \dots, x_t, y_1, \dots, y_t$" is meant to denote "almost-everywhere" with respect to the projection of γ onto $x_1, \dots, x_t, y_1, \dots, y_t$.

The ambient set throughout this article is $\mathbb{R}^N$, which we consider as a filtered space endowed with the canonical (i.e. coordinate) filtration $\left(\mathcal{F}_t\right)_{t=1}^N$. (Precisely $\mathcal{F}_t$ is the smallest σ-algebra on $\mathbb{R}^N$ such that the projection $\mathbb{R}^N \ni x \mapsto (x_1, \dots, x_t) \in \mathbb{R}^t$ onto the first t components is Borel-measurable, and so forth.)

We endow $\mathbb{R}^N$ with an ℓ^p-type product metric, namely

$$d(x,y) := d_p(x,y) := \sqrt[p]{\sum_{t=1}^{N} \underline{d}(x_i,y_i)^p}, \tag{2.1}$$

for some base metric $\underline{d}$ on $\mathbb{R}$ compatible with the usual topology and $p \in [1,\infty)$. We will be particularly interested in the cases where $\underline{d}$ is the usual distance or is a compatible bounded metric on $\mathbb{R}$. Notably, for most results one may substitute S^N for $\mathbb{R}^N$, where S is a Polish space, again endowing it with an ℓ^p-type product metric. Throughout this work, we fix $\underline{d}$, p and d as described.

The probability measures on the product space $\mathbb{R}^N \times \mathbb{R}^N$ with marginals μ and ν constitute the possible *transport plans* or *couplings* between the given marginals. We denote this set by

$$\Pi(\mu,\nu) = \left\{\gamma \in \mathcal{P}(\mathbb{R}^N \times \mathbb{R}^N) : \gamma \text{ has marginals } \mu \text{ and } \nu\right\}.$$

We often consider processes $X = \{X_t\}_{t=1}^N$, $Y = \{Y_t\}_{t=1}^N$, defined on some probability space. Each pair (X,Y) can be thought of as a coupling or – abusing notation slightly – as a transport plan upon identifying it with its law. For the sake of simplicity, being measurable with respect to a sigma algebra means to be equal to a correspondingly measurable function modulo a null set w.r.t. the measure relevant in the given context.

Definition 2.1. *A transport plan $\gamma \in \Pi(\mu,\nu)$ is called* bicausal *(between μ and ν) if the mappings*

$$\mathbb{R}^N \ni x \mapsto \gamma^x(B) \text{ and } \mathbb{R}^N \ni y \mapsto \gamma^y(B)$$

are $\mathcal{F}_t$-measurable for any $B \in \mathcal{F}_t$ and $t < N$. The set of all bicausal plans is denoted

$$\Pi_{bc}(\mu,\nu).$$

The product measure $\mu \otimes \nu$ is bi-causal, so $\Pi_{bc}(\mu,\nu)$ is non-empty. In terms of stochastic processes, a coupling is *bicausal* if

$$\mathbb{P}\big((Y_1,\dots,Y_t) \in B_t \mid X_1,\dots,X_N\big) = \mathbb{P}\big((Y_1,\dots,Y_t) \in B_t \mid X_1,\dots X_t\big) \text{ and}$$
$$\mathbb{P}\big((X_1,\dots,X_t) \in B_t \mid Y_1,\dots,Y_N\big) = \mathbb{P}\big((X_1,\dots,X_t) \in B_t \mid Y_1,\dots Y_t\big)$$

for all $t = 1,\dots,N$ and $B_t \subset \mathbb{R}^t$ Borel.

Testing whether a coupling or transport plan is bicausal reduces to a property of its transition kernel. Specifically we have the following characterization (see, e.g., [4])

Proposition 2.2. *The following are equivalent:*

1. *γ is a bicausal transport plan on $\mathbb{R}^N \times \mathbb{R}^N$ between the measures μ and ν.*
2. *The successive regular kernels $\bar\gamma$ of the decomposition*

$$\begin{aligned}&\gamma(\mathrm{d}x_1,\dots,\mathrm{d}x_N,\mathrm{d}y_1,\dots,\mathrm{d}y_N)\\ &\qquad = \bar\gamma(\mathrm{d}x_1,\mathrm{d}y_1)\gamma^{x_1,y_1}(\mathrm{d}x_2,\mathrm{d}y_2)\dots\gamma^{x_1,\dots,x_{N-1},y_1,\dots,y_{N-1}}(\mathrm{d}x_N,\mathrm{d}y_N)\end{aligned} \tag{2.2}$$

satisfy

$$\bar\gamma \in \Pi(p^1_*\mu, p^1_*\nu)$$

and further, for $t < N$ and γ-almost all $x_1,\dots,x_t,y_1,\dots,y_t$,

$$p^1_*\gamma^{x_1,\dots,x_t,y_1,\dots,y_t} = \mu^{x_1,\dots,x_t} \text{ and } p^2_*\gamma^{x_1,\dots,x_t,y_1,\dots,y_t} = \nu^{y_1,\dots,y_t}. \tag{2.3}$$

3. The nested distance

Following [18, 19, 20] we consider for μ, ν as above the *p-nested distance*, or simply *nested distance*, defined by

$$d_p^{\mathrm{nd}}(\mu,\nu) := \left(\inf_{\gamma\in\Pi_{bc}(\mu,\nu)} \iint d^p \mathrm{d}\gamma\right)^{1/p} = \left(\inf_{\gamma\in\Pi_{bc}(\mu,\nu)} \iint \left[\sum_{t=1}^{N} \underline{d}(x_t,y_t)^p\right] \mathrm{d}\gamma\right)^{1/p}, \tag{3.1}$$

obtaining a metric on the space

$$\mathcal{P}^p(\mathbb{R}^N) := \{\mu \in \mathcal{P}(\mathbb{R}^N) : \textstyle\int d(x,x_0)^p \mu(\mathrm{d}x) < \infty, \text{ some } x_0\},$$

in direct analogy with the classical p-Wasserstein space. As noted in [21], the nested distance (3.1) is best suited to "separate" μ and ν if their information structure differs. In particular, the authors show that empirical measures μ_n^{emp} of a multivariate measure μ with density never converge in nested distance (even though they do converge in Wasserstein distance); the essential point here is that each empirical measure μ_n^{emp} is roughly a tree with non-overlapping branches (commonly a *fan*) and therefore deterministic as soon as the first component is observed. From an information perspective this is radically different from μ. Arguably, this is a key property of the nested distance, and is its main distinctive characteristic and strength in comparison with the Wasserstein distance.

3.1. **Recursive computation.** A useful comment at this point is that the nested distance can be stated and computed recursively in a way which is comparable to Bellman equations: starting with $V_N^p := 0$ we define

$$V_t^p(x_1,\dots,x_t,y_1,\dots,y_t) := \inf_{\gamma^{t+1}\in\Pi(\mu^{x_1,\dots,x_t},\nu^{y_1,\dots,y_t})} \iint \begin{pmatrix} V_{t+1}^p(x_1,\dots,x_{t+1},y_1,\dots,y_{t+1}) \\ +\ \underline{d}(x_{t+1},y_{t+1})^p \end{pmatrix} \gamma^{t+1}(\mathrm{d}x_{t+1},\mathrm{d}y_{t+1}), \tag{3.2}$$

so that the nested distance is finally obtained in a backwards recursive way by

$$d_p^{\mathrm{nd}}(\mu,\nu)^p = \inf_{\gamma^1\in\Pi(p^1_*\mu,p^1_*\nu)} \iint \left(V_1^p(x_1,y_1) +\ \underline{d}(x_1,y_1)^p\right)\gamma^1(\mathrm{d}x_1,\mathrm{d}y_1). \tag{3.3}$$

3.2. **Comparison with weak topology.** The Wasserstein distance metrizes the weak topology on probability measures with suitably integrable moments. We recall that the weak topology (also called weak* or vague topology) is characterized by integration on bounded and continuous functions. It is thus natural to ask if there is a class of functions which characterizes the topology generated by the nested distance.

Proposition 3.1. *Let $N \geq 2$. There does* not *exist a family $\mathbb{F}$ of functions on $\mathbb{R}^N$ which determines convergence for d_p^{nd}. I.e., there is* no *family $\mathbb{F}$ so that*

$$d_p^{nd}(\mu_n,\mu)\to 0 \iff \int f\,\mathrm{d}\mu_n \to \int f\,\mathrm{d}\mu \text{ for all } f\in\mathbb{F}, \qquad (\mu_n)_{n=1}^\infty,\ \mu\in\mathcal{P}^p.$$

In fact, such a convergence determining family does not even exist if one restricts d_p^{nd} to distributions supported on a bounded region $[-K,K]^N\times[-K,K]^N$, $K\geq 0$.

Remark 3.2 (Separating evaluations)**.** *The nested distance was initially introduced with the intention to compare stochastic programs and the question answered by the preceding Proposition 3.1 was initially posed by Prof. Pflug. Indeed, Corollary 2 in* [19] *demonstrates that there are stochastic optimization programs with differing objective values whenever the nested distance differs.*

The separating objects are thus entire stochastic programs which, in view of the preceding Proposition 3.1, cannot be replaced by a set of functions on $\mathbb{R}^N$. This emphasizes further the intrinsic relation between stochastic programs, the nested distance, and the role of information.

Proof. Assume that such a family exists. Without loss of generality we can further assume that the integral of $f\in\mathbb{F}$ against all measures in $\mathcal{P}^p$ are well-defined. By considering $\delta_{(x_1^n,\dots,x_N^n)}$, which converge in nested distance to $\delta_{(x_1,\dots,x_N)}$ if their supports do in $\mathbb{R}^N$, we conclude that $\mathbb{F}\subset C(\mathbb{R}^N)$. Set

$$\mu_\varepsilon := \frac{1}{2}\left[\delta_{(\varepsilon,\dots,\varepsilon,1)}+\delta_{(-\varepsilon,\dots,-\varepsilon,-1)}\right] \quad\text{ and }\quad \mu := \frac{1}{2}\left[\delta_{(0,\dots,0,1)}+\delta_{(0,\dots,0,-1)}\right].$$

By continuity we find that $\int f\,\mathrm{d}(\mu_\varepsilon-\mu)\to 0$ as $\varepsilon\to 0$ for all $f\in\mathbb{F}$. Taking $\underline{d}$ to be the usual distance on $\mathbb{R}$ we find $d_p^{\mathrm{nd}}(\mu_\varepsilon,\mu)\geq 2^{p-1}$, for instance via (3.3). In general one sees that $d_p^{\mathrm{nd}}(\mu_\varepsilon,\mu)$ is bounded away of 0. Thus $\mathbb{F}$ cannot determine convergence in nested distance. □

We will see in Example 4.1 in the next section that d_p^{nd} is *not* complete. This further demonstrates how different the nested distance is from the usual Wasserstein distance.

Remark 3.3. *We stress that the metric results in Section 4 and the topological results in Section 5 are also applicable if we based the p-nested distance on an ℓ^q-type product norm in $\mathbb{R}$. Indeed, for each $q \in [1,\infty)$ we easily find c, $C>0$ s.t.*

$$c\, d(x,y) \le d_q(x,y) \le C\, d(x,y);$$

see (2.1) *for notation. In particular, if we base the p-nested distance* (3.1) *in terms of d_q instead of $d=d_p$, we obtain a strongly equivalent metric on $\mathcal{P}^p$ (with the same constants c and C). By the form of the metric d, we obtained an amenable expression for d_p^{nd}, as seen in the r.h.s. of* (3.1)*, which we would not have under d_q for $q \neq p$. For these reasons, we may and will continue to work with d_p^{nd} defined in terms of $d=d_p$ keeping in mind how the forthcoming results are trivially generalizable.*

4. Completeness and completion

The space $\mathcal{P}^p(\mathbb{R}^N)$, endowed with the p-Wasserstein distance is complete. This is not the case for the nested distance, as the following example reveals.

Example 4.1. *We observe that d_p^{nd} is not a complete metric as soon as the dimension N is greater or equal than 2. For the sake of the argument we take $N=2$, $\underline{d}$ the usual distance on $\mathbb{R}$, and consider $\mu_n = 1/2\{\delta_{(1/n,1)} + \delta_{(-1/n,-1)}\}$. One verifies that $d_p^{nd}(\mu_n,\mu_m) \le |1/n - 1/m|$, so the sequence is Cauchy. The only possible limit of this sequence is the limit based on the Wasserstein distance, that is $\mu = 1/2\{\delta_{(0,1)} + \delta_{(0,-1)}\}$. But in nested distance we have $d_p^{nd}(\mu_n,\mu) = (2^{p-1}+n^{-p})^{1/p} > 1$, in particular this sequence does not tend to zero.*

The distinguishing point is that μ is a real tree with coinciding states at the first stage, whereas the μ_n's are not. The nested distance is designed to capture this distinction, which is ignored by the Wasserstein distance.

So for $N>1$ the nested distance is not complete. To identify the completion of $\mathcal{P}^p(\mathbb{R}^N)$ with respect to the p-nested distance, we consider the *nested distributions* introduced in [18].

Definition 4.2. *Consider the sequence of metric spaces*

$$\begin{aligned}
R_{N:N} &:= (\mathbb{R}, d_{(N:N)}), \text{ with } d_{(N:N)} = \underline{d} = [\underline{d}^p]^{1/p}\\
R_{N-1:N} &:= \big(\mathbb{R}\times\mathcal{P}^p(R_{N:N}), d_{(N-1:N)}\big), \text{ with } d_{(N-1:N)} = \Big[\underline{d}^p + W^p_{d_{(N:N)},p}\Big]^{1/p}\\
&\;\;\vdots\\
R_{1:N} &:= \big((\mathbb{R}\times\mathcal{P}^p(R_{2:N})), d_{(1:N)}\big), \text{ with } d_{(1:N)} = \Big[\underline{d}^p + W^p_{d_{(2:N)},p}\Big]^{1/p},
\end{aligned}$$

where at each stage t, the space $\mathcal{P}^p(R_{t:N})$ is endowed with the p-Wasserstein distance with respect to the metric $d_{(t:N)}$ on $R_{t:N}$, which we denote $W_{d_{(t:N)},p}$. The set of nested distributions *(of depth N) with p-th moment is defined as $\mathcal{P}^p(R_{1:N})$.*

Each of the spaces $R_{t:N}$ $(t=1,\dots N)$ is a Polish space. Indeed, a complete metric is given explicitly, and the spaces are separable since $\mathcal{P}(R)$ is complete and separable whenever (R,ρ) is complete and separable (cf. [9]). We endow $\mathcal{P}^p(R_{1:N})$ with the complete metric $W_{d_{(1:N)},p}$.

Example 4.3. *When $N=2$, we have that $R_{1:2} = \mathbb{R}\times\mathcal{P}^p(\mathbb{R})$ and for $P,Q\in\mathcal{P}^p(R_{1:2})$ the distance is*

$$W_{d_{(1:2)},p}(P,Q) = \big\{\inf_{\Gamma\in\Pi(P,Q)} \textstyle\iint \big(\underline{d}(x,y)^p + W_p^p(\mu,\nu)\big)\Gamma(\mathrm{d}x,\mathrm{d}\mu,\mathrm{d}y,\mathrm{d}\nu)\big\}^{1/p} \tag{4.1}$$

with W_p the classical p-Wasserstein distance for measures on the line and w.r.t the metric $\underline{d}$. The formulation (4.1) *notably exactly corresponds to the recursive descriptions* (3.2) *and* (3.3).

4.1. **Embedding.** We demonstrate that the nested distributions of depth N introduced in Definition 4.2 extend the notion of probability measures in $\mathbb{R}^N$ in a metrically meaningful way. Let us introduce the following function, already present in [18], which associates $\mu\in\mathcal{P}^p(\mathbb{R}^N)$ with the measure $I[\mu]\in\mathcal{P}^p(R_{1:N})$ given by

$$I[\mu] := Ł\left(X_1\,,\, Ł^{X_1}\Big(X_2\,,\,\cdots\,,\,Ł^{X_{1:N-2}}\big(X_{N-1}\,,\,Ł^{X_{1:N-1}}(X_N)\big)\Big)\right), \tag{4.2}$$

where $(X_1,\dots,X_N)$ is a vector with law μ. We used the shorthand $\mathrm{L}^{X_{1:k}}$ for the *conditional law* given $(X_1,\dots,X_k)$ (and no superscript indicates unconditional law).

Theorem 4.4. *Let $d=d_p$. Then the classical Wasserstein distance of nested distributions extends the nested distance of classical distributions. More precisely, the mapping I defined in* (4.2) *embeds the metric space $(\mathcal{P}^p(\mathbb{R}^N), d_p^{nd})$ defined via* (3.1) isometrically *into the separable complete metric space $(\mathcal{P}^p(R_{1:N}), W_{d_{(1:N)},p})$. In particular $(\mathcal{P}^p(\mathbb{R}^N), d_p^{nd})$ is separable.*

Proof. It is enough to consider $N=2$. For a probability measure μ on $\mathbb{R}^2$ consider its disintegration measure

$$\mu(A\times B)=\textstyle\int_A \mu^{x_1}(B)p^1_*\mu(\mathrm{d}x_1),$$

where p^1 is the projection onto the first coordinate. An embedding of μ in the space $\mathcal{P}\big(\mathbb{R}\times\mathcal{P}(\mathbb{R})\big)$ is given by the probability measure generated uniquely by (here A,B are Borel sets of $\mathbb{R}$ and $\mathcal{P}(\mathbb{R})$ respectively)

$$I[\mu](A\times B):=\mu\big(A\cap T_\mu^{-1}(B)\big)=p^1_*\mu\big(A\cap T_\mu^{-1}(B)\big),$$

where T_μ is the Borel measurable function

$$\begin{aligned} T_\mu:\mathbb{R}&\to\mathcal{P}(\mathbb{R})\\ x_1&\mapsto \mu^{x_1}(\mathrm{d}x_2) \end{aligned}$$

In this way we find that $I[\mu]$ is the μ-law of $x_1\mapsto(x_1,\mu^{x_1}(\mathrm{d}x_2))$. For $\mu\in\mathcal{P}^p(\mathbb{R}^2)$ we also have

$$\begin{aligned}\int\big\{\underline{d}(x,0)^p+W_p^p(\nu,\delta_0)\big\}\,I[\mu](\mathrm{d}x,\mathrm{d}\nu)&=\int\big\{\underline{d}(x_1,0)^p+W_p^p(\mu^{x_1},\delta_0)\big\}\,p^1_*\mu(\mathrm{d}x_1)\\ &=\int\{\underline{d}(x_1,0)^p+\underline{d}(x_2,0)^p\}\,\mu(\mathrm{d}x_1,\mathrm{d}x_2)<\infty\end{aligned}$$

and thus $I[\mu]\in\mathcal{P}^p(R_{1:2})$.

We now observe that the embedding $\mu\mapsto I[\mu]$ is actually an isometry between $(\mathcal{P}^p(\mathbb{R}^N), d_p^{\mathrm{nd}})$ and $(\mathcal{P}^p(R_{1:N}), W_{d_{(1:N)},p})$. To this end, first note that every coupling between $I[\mu]$ and $I[\nu]$ (i.e., every $\Gamma\in\Pi(I[\mu],I[\nu])$) is of the form $\bar\gamma(\mathrm{d}x_1,\mathrm{d}y_1)\delta_{T_\mu(x_1)}(\mathrm{d}M)\delta_{T_\nu(y_1)}(\mathrm{d}N)$ for some $\bar\gamma\in\Pi(p^1_*\mu,p^1_*\nu)$ and vice-versa. Hence from (4.1) and (3.2) we have that

$$\begin{aligned}W_{d_{(1:2)},p}\big(I[\mu],I[\nu]\big)^p&=\inf_{\bar\gamma\in\Pi(p^1_*\mu,p^1_*\nu)}\int\big\{\underline{d}(x_1,y_1)^p+W_p^p(\mu^{x_1},\nu^{y_1})\big\}\,\bar\gamma(\mathrm{d}x_1,\mathrm{d}y_1)\\ &=d_p^{\mathrm{nd}}(\mu,\nu)^p,\end{aligned}$$

by (3.3), and hence the isometry. Finally, since the image of I is a subspace of the separable metric space $(\mathcal{P}^p(R_{1:N}), W_{d_{(1:N)},p})$, it is separable itself. We conclude that $(\mathcal{P}^p(\mathbb{R}^N), d_p^{\mathrm{nd}})$ is separable too. □

Remark 4.5. *It follows from the preceding arguments as well that the embedding I is onto if and only if $N=1$.*

4.2. **Completion.** We now identify the completion of $(\mathcal{P}^p(\mathbb{R}^N), d_p^{\mathrm{nd}})$. (Recall that the completion of a metric space is unique up to isomorphism.) This result is unexpected and provides a solid link between these two previously separate mathematical objects.

Theorem 4.6. *The space $\big(\mathcal{P}^p(R_{1:N}), W_{d_{(1:N)},p}\big)$ is the completion of $\big(\mathcal{P}^p(\mathbb{R}^N), d_p^{nd}\big)$.*

Proof. We need to provide an isometry J from $(\mathcal{P}^p(\mathbb{R}^N), d_p^{\mathrm{nd}})$ into $(\mathcal{P}^p(R_{1:N}), W_{d_{(1:N)},p})$ whose range is dense. We shall prove that I defined in (4.2) does this task. This can be done for arbitrary N at a notational, while already the case $N=2$ is representative of the general situation. We thus assume $N=2$ in what follows.

The set of convex combinations of Dirac measures is dense in $\mathcal{P}^p(R_{1:N})$ w.r.t. the metric $W_{d_{(1:N)},p}$. This is actually true for any Wasserstein metric (cf. [9]) and thus particularly for $W_{d_{(1:2)},p}$, which in itself is a Wasserstein metric (see also Example 4.3 for concreteness). So it suffices to prove that convex combinations of Dirac measures lie in the closure of the range of I.

Let $A:=(a_1,\dots,a_k)$ be a k-tuple of points in $\mathbb{R}$ and $m_1,\dots,m_k$ be measures on the line with finite p-th moment. Given weighs $\{\lambda_i\}_{i=1}^k$ we are interested in the measure

$$P(\mathrm{d}x,\mathrm{d}m)=\sum\lambda_i\delta_{(a_i,m_i)}(\mathrm{d}x,\mathrm{d}m),$$

over $R_{1:2}$. Now we take any sequence $A^n := \{a_1^n, \ldots, a_k^n\}$ such that componentwise $A^n \to A$ as $n \to \infty$ and, for each n fixed, all coordinates of A^n are distinct. We now define $\mu_n \in \mathcal{P}^p(\mathbb{R}^2)$ as the measure whose first marginal is $\sum \lambda_j \delta_{a_j^n}$ and such that $\mu_n(\mathrm{d}x_2|x_1 = a_j^n) = m_j(\mathrm{d}x_2)$. It is elementary, and this is the main point of having made the a_j^n's distinct for a fixed n, that

$$I[\mu_n] = \sum \lambda_j \delta_{(a_j^n, m_j)}.$$

Consequently we get that $I[\mu_n] \to P$ with respect to $W_{d_{(1:2)},p}$ when $n \to \infty$, as desired.

□

At this point we ask,

$$\text{Can } (\mathcal{P}^p(\mathbb{R}^N), d_p^{\mathrm{nd}}) \text{ still be Polish, for } N \geq 2?.$$

Just as the interval $(0,1)$ is a Polish subspace of $[0,1]$, which is nevertheless incomplete w.r.t. the usual distance on $[0,1]$, neither Example 4.1 nor Theorem 4.6 contribute anything to this question. We explore this in the following section.

5. The information topology / weak nested topology

We introduce the space $\mathcal{P}(R_{1:N})$ just as we did for $\mathcal{P}^p(R_{1:N})$, but now denoting $R_{t-1:N} := \mathbb{R} \times \mathcal{P}(R_{t:N})$ at each step of the recursive definition, and equipping $R_{t-1:N}$ with the product topology of Euclidean distance in the first component and the usual weak topology in the second one. Doing so, we conclude that $\mathcal{P}(R_{1:N})$ is a Polish space of measures on the likewise Polish space $R_{1:N}$. Inspired by the isometric embedding in Theorem 4.4, which we denoted I in (4.2), a mapping $I : \mathcal{P}(\mathbb{R}^N) \to \mathcal{P}(R_{1:N})$ can be obtained by direct generalization.

Definition 5.1. *We say that a net $\{\mu_\alpha\}_\alpha$ in $\mathcal{P}(\mathbb{R}^N)$* weakly nested *converges to $\mu \in \mathcal{P}(\mathbb{R}^N)$, if and only if $I[\mu_\alpha]$ converges weakly in $\mathcal{P}(R_{1:N})$ to $I[\mu]$. We call the corresponding topology* weak nested topology*.*

Thus the weak nested topology is simply the initial topology for the map I.

Remark 5.2. *Suppose that X is Polish and that ρ is a compatible complete metric which is* bounded*. Then the corresponding p-Wasserstein topology is precisely the weak topology. Likewise we obtain that the weak nested topology is generated by the nested distance d_p^{nd} as soon as we choose $\underline{d}$ as a compatible bounded metric for the usual topology on $\mathbb{R}$. For instance, we may take the metric $\underline{d}(a,b) = |a-b| \wedge 1$ so*

$$d(x,y) = \sum_{t=1}^{N} |x_t - y_t| \wedge 1.$$

In this way we obtain that the weak nested topology coincides with a p-nested topology of the form we have already treated.

Although there are more direct ways to prove it, the previous remark implies the following:

Lemma 5.3. *The weak nested topology is separable and metrizable.*

5.1. **Comparison with an existing concept.** Definition 5.1 is related to the so-called *topology of information* which was introduced in [13] for the purpose of sequential decision problems and equilibria (see also [6] for a recent update). In our setting, this topology is defined as the initial topology for the following maps on $\mathcal{P}(\mathbb{R}^N)$:

$$\begin{aligned}
&\mu \mapsto \mu \quad \in \mathcal{P}(\mathbb{R}^N)\\
&\mu \mapsto [x_1 \mapsto (x_1, \mu^{x_1}(\mathrm{d}x_2, \ldots, \mathrm{d}x_N))]_*\mu \quad \in \mathcal{P}(\mathbb{R} \times \mathcal{P}(\mathbb{R}^{N-1}))\\
&\mu \mapsto [(x_1,x_2) \mapsto (x_1, x_2, \mu^{x_1,x_2}(\mathrm{d}x_3, \ldots, \mathrm{d}x_N))]_*\mu \quad \in \mathcal{P}(\mathbb{R}^2 \times \mathcal{P}(\mathbb{R}^{N-2}))\\
&\quad \vdots\\
&\mu \mapsto [(x_1,\ldots,x_{N-1}) \mapsto (x_1, \ldots, x_{N-1}, \mu^{x_1,\ldots,x_{N-1}}(\mathrm{d}x_N))]_*\mu \quad \in \mathcal{P}(\mathbb{R}^{N-1} \times \mathcal{P}(\mathbb{R})),
\end{aligned}$$

where the range spaces are endowed with the usual weak topologies.

For $N = 1, 2$ this topology obviously coincides with the weak nested one of Definition 5.1. As a matter of fact, this is always the case. We show the argument for $N = 3$:

Let A, B continuous bounded function on $\mathbb{R} \times \mathcal{P}(\mathbb{R}^2)$ and $\mathbb{R}^2 \times \mathcal{P}(\mathbb{R})$, respectively. We denote by m, M generic elements in $\mathcal{P}(\mathbb{R})$ and $\mathcal{P}(\mathbb{R} \times \mathcal{P}(R))$, respectively. Then

$$\mathbb{R} \times \mathcal{P}(\mathbb{R} \times \mathcal{P}(R)) \ni (x_1, M) \mapsto \bar{A}(x_1, M) := A\left(x_1, \int_{\mathbb{R}\times\mathcal{P}(\mathbb{R})} m\, M(\mathrm{d}x_2, \mathrm{d}m)\right),$$
$$\mathbb{R} \times \mathcal{P}(\mathbb{R} \times \mathcal{P}(R)) \ni (x_1, M) \mapsto \bar{B}(x_1, M) := \int_{\mathbb{R}\times\mathcal{P}(\mathbb{R})} B(x_1, x_2, m) M(\mathrm{d}x_2, \mathrm{d}m),$$

are seen to be continuous bounded functions too, thus they are suitable test functions for weak nested convergence (since $R_{1:3} = \mathbb{R} \times \mathcal{P}(\mathbb{R} \times \mathcal{P}(R))$ precisely). One then easily verifies that

$$\int \bar{A}(x_1, M) I[\mu](\mathrm{d}x_1, \mathrm{d}M) = \int_{\mathbb{R}} A(x_1, \mu^{x_1}(\mathrm{d}x_2, \mathrm{d}x_3))\mu(\mathrm{d}x_1),$$
$$\int \bar{B}(x_1, M) I[\mu](\mathrm{d}x_1, \mathrm{d}M) = \int_{\mathbb{R}^2} B(x_1, x_2, \mu^{x_1,x_2}(\mathrm{d}x_3))\mu(\mathrm{d}x_1, \mathrm{d}x_2),$$

so convergence of the l.h.s (guaranteed by weak nested convergence) implies that of the r.h.s. which then implies convergence in information topology. Thus the weak nested topology is stronger than the topology of information. For the converse, recall first that

$$\mu \in \mathcal{P}(\mathbb{R}^3) \mapsto I[\mu] = Ł(X_1\, , Ł^{X_1}(X_2\, , Ł^{X_1,X_2}(X_3))),$$

where (X_1, X_2, X_3) is distributed according to μ. If we denote

$$m \in \mathcal{P}(\mathbb{R}^2) \mapsto T[m] := [x \mapsto (x, m^x(\mathrm{d}y))]_* m \in \mathcal{P}(\mathbb{R} \times \mathcal{P}(\mathbb{R})),$$
$$(x, m) \in \mathbb{R} \times \mathcal{P}(\mathbb{R}^2) \mapsto L(x, m) := (x, T[m]) \in \mathbb{R} \times \mathcal{P}(\mathbb{R} \times \mathcal{P}(\mathbb{R})),$$
$$p \in \mathcal{P}(\mathbb{R}^3) \mapsto \phi(p) := [x \mapsto (x, p^x(\mathrm{d}y, \mathrm{d}z))]_* p \in \mathcal{P}(\mathbb{R} \times \mathcal{P}(\mathbb{R}^2)),$$

one finds that $I[\mu] = L_*\phi(\mu)$. By definition ϕ is continuous in topology of information. The key now is [13, Lemma 7], from which ϕ is also continuous if on the range space, $\mathcal{P}(\mathbb{R} \times \mathcal{P}(\mathbb{R}^2))$, we endow $\mathcal{P}(\mathbb{R}^2)$ with the information topology again. Since L is continuous when the domain is given this topology, we finally conclude that $L_*\phi(\mu)$ is continuous in information topology, making the latter stronger than the weak nested one.

Because we are inspired by the nested distance of Pflug and Pichler, and due to the observation in Remark 5.2 that the weak nested topology (so a fotriori the information topology) is a nested distance topology, we shall use the term “weak nested topology” instead of “information topology” in the following. The results to come are also new in the setting of [13, 6].

5.2. **A closer look at the weak nested topology.** We will establish that the weak nested topology (and actually the nested distance topologies) is Polish. Although we deem this interesting per-se, we hope this can find future applications too.

We recall that a set of a topological spaces is a G_δ if it is the countable intersection of open sets. Recall also that every separable metrizable space is homeomorphic to a subspace of the Hilbert cube $[0,1]^{\mathbb{N}}$, the latter equipped with the product topology; see [14, Theorem 4.14]. A compatible metric on the Hilbert cube is given by

$$D\big((x_n), (y_n)\big) := \sum_{n=1} 2^{-n}|x_n - y_n|.$$

Lemma 5.4. *Let $m \in \mathcal{P}(X \times Y)$ with X Polish and (Y, ρ) a separable metric space. Denote $\iota : Y \to [0,1]^{\mathbb{N}}$ the embedding of Y into the Hilbert cube. Then the following are equivalent:*

(1) $m\big(\mathit{Graph}(f)\big) = 1$ *for* $f : X \to Y$ *Borel;*
(2) $\inf\left\{\int_{X\times Y} \rho\big(f(x), y\big) m(\mathrm{d}x, \mathrm{d}y) :\ f : X \to Y \text{ Borel}\right\} = 0;$
(3) $\inf\left\{\int_{X\times Y} D\big(F(x), \iota(y)\big) m(\mathrm{d}x, \mathrm{d}y) :\ F : X \to [0,1]^{\mathbb{N}} \text{ Borel}\right\} = 0;$
(4) $\inf\left\{\int_{X\times Y} D\big(F(x), \iota(y)\big) m(\mathrm{d}x, \mathrm{d}y) :\ F : X \to [0,1]^{\mathbb{N}} \text{ continuous}\right\} = 0.$

Proof. Clearly $1 \implies 2 \implies 3$. Denote μ the first marginal of m. Given $F : X \to [0,1]^{\mathbb{N}}$ Borel, $F(x) = (F_n(x))_n$, we can approximate it in $L^1(X, \mu; [0,1]^{\mathbb{N}})$ by continuous functions. This follows since coordinate-wise we can approximate $F_n \in L^1(X, \mu; [0,1])$ by continuous functions. So also $3 \implies 4$.

To establish $4 \implies 1$ let $\{F_n\}$ be a sequence of continuous functions approximating the infimum in 4, and denote $G_n(x) := \int D\big(F_n(x), \iota(y)\big) m^x(\mathrm{d}y)$ so G_n is Borel, non-negative, and $\|G_n\|_{L^1(X,\mu;\mathbb{R})} \to 0$ by definition. It follows that $G_n \to 0$ in $L^1(X, \mu; \mathbb{R})$ so up to a subsequence $G_n(x) \to 0$ for μ-a.e. x. From now on we work on such a full measure set, on which we can further assume that $m^x \in \mathcal{P}(Y)$. Assume that we had that $|supp(m^x)| > 1$. Then there would exist disjoint compact sets K_x^1, $K_x^2 \subset Y$ with $M_x = \min\{m^x(K_x^1), m^x(K_x^2)\} > 0$. Obviously $\iota(K_x^1), \iota(K_x^2)$ are also disjoint compact sets, so $D_x := D\big(\iota(K_x^1), \iota(K_x^2)\big) > 0$. By the triangle inequality, $\max\{D\big(F_n(x), \iota(K_x^1)\big), D\big(F_n(x), \iota(K_x^2)\big)\} \geq D_x/2$, thus $G_n(x) \geq M_x D_x/2$, yielding a contradiction. We conclude that μ-a.s. $|supp(m^x)| = 1$ and therefore we must have $\iota\big(f(x)\big) := \lim_n f_n(x)$ exists, for some $f : X \to Y$ Borel. Thus $m^x(\mathrm{d}y) = \delta_{f(x)}(\mathrm{d}y), \mu - a.s.$, which proves 1.

□

Observe that it is crucial for point 4 in Lemma 5.4 that we embedded Y in the Hilbert cube. Indeed, it X is connected and Y discrete, the only continuous functions $f : X \to Y$ are the constants. We now present a result which is interesting in its own:

Proposition 5.5. *Let X and Y be Polish spaces. Then*

$$S := \{m \in \mathcal{P}(X \times Y) : m\big(Graph(f)\big) = 1,\ some\ Borel\ f : X \to Y\},$$

with the relative topology inherited from $\mathcal{P}(X \times Y)$, is Polish too.

Proof. Let ρ be a compatible metric for Y, which we may assume bounded. By Lemma 5.4 we have

$$S = \bigcap_{n\in\mathbb{N}} \bigcup_{F:X\to[0,1]^{\mathbb{N}} \text{ continuous}} \{m \in \mathcal{P}(X \times Y) : \int D\big(F(x), \iota(y)\big) m(\mathrm{d}x, \mathrm{d}y) < 1/n\}, \tag{5.1}$$

where $\iota : Y \to [0, 1]^{\mathbb{N}}$ is an embedding. Since $(x, y) \mapsto D\big(F(x), \iota(y)\big)$ is continuous bounded if F is continuous, the set in curly brackets is open in the weak topology. Thus the union of these is open too, and we get that S is a G_δ subset. We conclude by employing [14, Theorem 3.11].

□

Theorem 5.6. *The weak nested topology on $\mathcal{P}(\mathbb{R}^N)$ is Polish.*

Proof. For $N = 2$ we have $\mathcal{P}(R_{1:2}) = \mathcal{P}(\mathbb{R} \times \mathcal{P}(\mathbb{R}))$ and, by definition, $\mathcal{P}(\mathbb{R}^2)$ equipped with the weak nested topology is homeomorphic to $I[\mathcal{P}(\mathbb{R}^2)]$ equipped with the relative topology inherited from $\mathcal{P}(R_{1:2})$. We have

$$I[\mathcal{P}(\mathbb{R}^2)] = \big\{P \in \mathcal{P}(\mathbb{R} \times \mathcal{P}(\mathbb{R})) : P\big(\mathrm{Graph}(f)\big) = 1, \text{ some Borel } f : \mathbb{R} \to \mathcal{P}(\mathbb{R})\big\}.$$

To wit, if $P \in I[\mathcal{P}(\mathbb{R}^2)]$, then by definition of the embedding I we have $P = (id, T)_*(p^1_*\mu)$ for some $\mu \in \mathcal{P}(\mathbb{R}^2)$ and $T(x) = \mu^x$ (see the proof of Theorem 4.4 too). Then taking $f = T$ we get that P belongs to the right hand side above. Conversely, given P in the right hand side, we denote by μ_1 its first marginal and define $\mu^x(dy) := \delta_{f(x)}(dy)$. The measure $\mu(dx, dy) := \mu_1(dx)\mu^x(dy) \in \mathcal{P}(\mathbb{R}^2)$ satisfies $I[\mu] = P$.

By Proposition 5.5 we conclude that $I[\mathcal{P}(\mathbb{R}^2)]$ is Polish, and then so is $\mathcal{P}(\mathbb{R}^2)$, as desired. The case for general N is identical; one observes by reverse induction that if $\mathcal{P}(R_{t:N})$ is Polish, then so is $\mathcal{P}(R_{t-1:N})$ using the above arguments.

□

One can also prove Theorem 5.6 by a suitable modification of Lemma 5.4, with the advantage of obtaining a complete compatible metric for the Polish topology, and this we do next. With similar arguments, a more involved complete metric can also be found via Proposition 5.5, as the proof will reveal. For simplicity of notation we just consider $N = 2$ here:

Corollary 5.7. *Let ρ be a bounded metric compatible with the weak topology on $\mathcal{P}(\mathbb{R})$, and d^w a complete metric compatible with the weak topology on $\mathcal{P}(\mathbb{R} \times \mathcal{P}(\mathbb{R}))$. Then the weak nested topology on $\mathcal{P}(\mathbb{R}^2)$ is generated by the complete metric*

$$d^{wnt}(P, Q) := d^w(I[P], I[Q]) + \sum_{n\in\mathbb{N}} 2^{-n} \wedge \left| \frac{1}{d^w(I[P], A_n)} - \frac{1}{d^w(I[Q], A_n)} \right|, \tag{5.2}$$

where $A_n := \{m \in \mathcal{P}(\mathbb{R}\times\mathcal{P}(\mathbb{R})) : \int \rho(F(x),y)m(dx,dy) \geq 1/n, \forall\, F:\mathbb{R}\to\mathcal{P}(\mathbb{R})$ *continuous*$\}$, *with the embedding* I *as in* (4.2), *and*

$$d^w(\cdot, A_n) := \inf_{m\in A_n} d^w(\cdot, m).$$

Proof. We first observe that for Lemma 5.4 and $Y=\mathcal{P}(\mathbb{R})$, we can by-pass the embedding into the Hilbert cube. One way to do this, is to follow the "Tietze extension" argument in the proof of [10, Proposition C.1], establishing the equivalence of (1) and (4) in Lemma 5.4 where now the continuous functions go from $X=\mathbb{R}$ to $\mathcal{P}(\mathbb{R})$. We can thus write (5.1), in the case $Y=\mathcal{P}(\mathbb{R})$, without the embedding ι. Using this, and following the proof of [14, Theorem 311], we find a compatible complete metric for $I[\mathcal{P}(\mathbb{R}^2)]$ with the relative topology inherited from $\mathcal{P}(\mathbb{R}\times\mathcal{P}(\mathbb{R}))$, via:

$$I[\mathcal{P}(\mathbb{R}^2)] \ni \bar P, \bar Q \mapsto d^w(\bar P,\bar Q) + \textstyle\sum_n 2^{-n}\wedge \left| d^w(\bar P, A_n)^{-1} - d^w(\bar Q, A_n)^{-1}\right|.$$

This is then transformed into a complete metric for $\mathcal{P}(\mathbb{R}^2)$ via the homeomorphism I, yielding (5.2). $\square$

Remark 5.8. *Notice that Example 4.1 shows that the weak nested topology is strictly stronger than the weak topology for $N\geq 2$. In such case, it also shows that even if a sequence of measures has their support contained in a common compact, there need not exist a convergent subsequence, unlike in the weak topology. It could be interesting, and non-trivial, to characterize the relatively compact sets of the weak nested topology.*

Analogous considerations show that $\mathcal{P}^p(R_{1:N})$ with the p-nested distance is Polish as well. Having established the completion and the Polish character of the p-nested distance, it remains an open question, whether there is a more amenable compatible complete metric that the one found in Corollary 5.7. Instead, we now investigate a different and appealing notion of distance, which we compare to the nested distance.

6. The Knothe–Rosenblatt distance

Perhaps the most eminent of bicausal plans (i.e., those participating in the determination of the nested distance) is the so-called increasing Knothe–Rosenblatt rearrangement, which is also known as quantile transform or increasing triangular transform in the literature. See, e.g., [4] and references therein. Let us introduce some useful notation first. By $F_\eta(\cdot)$ we denote the distribution function of a probability measure η on the line (we denote F_{ν_1} the distribution of $p^1_*\nu$) and by $F^{-1}_\eta(\cdot)$ its left-continuous generalized inverse, i.e., $F^{-1}_\eta(u)=\inf\{y : F_\eta(y)\geq u\}$. The (increasing N-dimensional) *Knothe–Rosenblatt* rearrangement of μ and ν is defined as the law π of the random vector $(X^*_1,\dots,Y^*_N,X^*_1,\dots,Y^*_N)$, where

$$\begin{aligned} X^*_1 &= F^{-1}_{\mu_1}(U_1), & Y^*_1 &= F^{-1}_{\nu_1}(U_1), && \text{and inductively}\\ X^*_t &= F^{-1}_{\mu^{X^*_1,\dots,X^*_{t-1}}}(U_t), & Y^*_t &= F^{-1}_{\nu^{Y^*_1,\dots,Y^*_{t-1}}}(U_t), && \text{for } t=2,\dots,N, \end{aligned} \tag{6.1}$$

for $U_1,\dots,U_N$ independent and standard uniformly distributed random variables. Additionally, if μ-a.s. all the conditional distributions of μ are atomless (e.g., if μ has a density), then this rearrangement is induced by the map

$$(x_1,\dots,x_N)\mapsto T(x_1,\dots,x_N) := (T^1(x_1), T^2(x_2|x_1),\dots,T^N(x_N|\,x_1,\dots x_{N-1})),$$

where

$$\begin{aligned} T^1(x_1) &:= F^{-1}_{\nu_1}\circ F_{\mu_1}(x_1),\\ T^t(x_t|\,x_1,\dots,x_{t-1}) &:= F^{-1}_{\nu^{T^1(x_1),\dots,T^{t-1}(x_{t-1}|\,x_1,\dots,x_{t-2})}}\circ F_{\mu^{x_1,\dots,x_{t-1}}}(x_t), \quad t\geq 2. \end{aligned} \tag{6.2}$$

In this section we reserve the letters π and T for this rearrangement (map) and omit its dependence on (μ,ν), which is clear from the context. Let us now define a functional on $\mathcal{P}^p(\mathbb{R}^N)\times\mathcal{P}^p(\mathbb{R}^N)$ which we compare with the nested distance in Section 6.1 below.

Definition 6.1. *The* Knothe–Rosenblatt distance *of order p (in short KR distance) is defined by*

$$d^{KR}_p(\mu,\nu) := \left(\iint d^p\,\mathrm{d}\pi\right)^{1/p}, \tag{6.3}$$

where π is the Knothe–Rosenblatt rearrangement of μ and ν.

Lemma 6.2. *The KR distance d^{KR}_p is a metric on $\mathcal{P}^p(\mathbb{R}^N)$ for any $1\leq p<\infty$.*

Proof. Since both d^p and the Knothe–Rosenblatt rearrangement are symmetric, we see that d_p^{KR} is symmetric. Obviously this distance is non-negative and vanishes exactly when $\mu = \nu$, since one dimensional distributions and conditional distributions fully encode a measure. Finally observe that $d_p^{KR}(\mu,\nu) = (E[d(X^*,Y^*)^p])^{1/p}$, where X^*, Y^* are as in (6.1). Given $\eta \in \mathcal{P}^p(\mathbb{R}^N)$ and constructing Z^* as in (6.1) so that it is distributed like η, the triangle inequality follows from $(E[d(X^*,Y^*)^p])^{1/p} \leq (E[d(X^*,Z^*)^p])^{1/p} + (E[d(Z^*,Y^*)^p])^{1/p}$. □

For its simplicity, and the fact that the Knothe–Rosenblatt rearrangement has already found startling applications all over mathematics, one is tempted to explore the connection between the KR and the nested distances. Furthermore, by the optimality properties of the increasing rearrangement on the line, it is clear that $d_p^{nd} = d_p^{KR}$ for $N = 1$ and e.g. $d(x,y) = |x-y|$, in which case both metrics coincide with the usual p-Wasserstein metric on $\mathcal{P}(\mathbb{R})$. Can it be that d_p^{nd} is always comparable/similar, if not equal, to d_p^{KR}?. We examine this now.

6.1. **Relationship with the nested distance.** By definition we have that

$$d_p^{\mathrm{nd}}(\mu,\nu) \leq d_p^{KR}(\mu,\nu). \tag{6.4}$$

We ask,

(1) Is there a constant $C > 0$ such that $d_p^{KR}(\mu,\nu) \leq C d_p^{\mathrm{nd}}(\mu,\nu)$ holds, i.e., are the two metrics strongly equivalent?
(2) Is it the case that $d_p^{\mathrm{nd}}(\mu_n,\mu) \to 0$ implies $d_p^{KR}(\mu_n,\mu) \to 0$, i.e., are the two metrics topologically equivalent?

From now we specialize to the case $N = 2$ but the situation is the same for all $N > 1$. The next counterexample shows that the answer to question 1 is negative:

Example 6.3. *Let $\mu_n := 1/2[\delta_{(1/n,n/2)} + \delta_{(-1/n,-n/2)}]$ and $\nu_n := 1/2[\delta_{(1/n,-n/2)} + \delta_{(-1/n,n/2)}]$ and take $\underline{d}$ the usual distance in $\mathbb{R}$. Observe that any transport plan is bicausal in this setting. Thus, we bound $d_p^{nd}(\mu_n,\nu_n)$ from above by the value of the decreasing Knotte-Rosenblatt map (associating $\{1/n, n/2\}$ to $\{-1/n, n/2\}$ and $\{-1/n,-n/2\}$ to $\{1/n,-n/2\}$)*

$$d_p^{nd}(\mu_n,\nu_n) \leq 2/n,$$

whereas for the usual (increasing) Knotte-Rosenblatt map we get

$$d_p^{KR}(\mu_n,\nu_n) = n,$$

and letting $n \to \infty$ we see that 1 cannot hold.

Actually, even 2 fails, as the following counterexample demonstrates:

Example 6.4. *Consider the measure $P := \lambda \otimes \lambda$ on $\Omega := [0,1]^2$ and the random variables (i.e., the two-stage stochastic processes)*

$$Z_n := \begin{cases} (0,\ u_2) & \text{if } u_1 \leq \frac{1}{2}, \\ (\frac{1}{n},\ 1+u_2) & \text{if } u_1 > \frac{1}{2} \end{cases}$$

for $n = 1,2,\dots$ and

$$Z_\infty := \begin{cases} (0,\ u_2) & \text{if } u_1 \leq \frac{1}{2}, \\ (0,\ 1+u_2) & \text{if } u_1 > \frac{1}{2} \end{cases}$$

together with their image measures

$$\mu^{(n)} := P \circ Z_n^{-1} \text{ and } \mu := P \circ Z_\infty^{-1}.$$

It is evident that we have convergence in nested distance, $d_p^{nd}(\mu_n,\mu) \to 0$ as $n \to \infty$, as we may employ the conditional transport maps

$$T(x|\,U_1) = x$$

on the second stage (i.e., $T(u_2|\,u_1 \geq \frac{1}{2}) = u_2$ and $T(1-u_2|\,u_1 < \frac{1}{2}) = 1-u_2$).

To compute the Knothe–Rosenblatt distance we associate $\mu^{(n)}$ with X^* in (6.1) *and μ with Y^*. Then*

$$X_1^* = \begin{cases} 0 & \text{if } U_1 \leq \frac{1}{2}, \\ \frac{1}{n} & \text{if } U_1 > \frac{1}{2} \end{cases} \quad \text{and} \quad Y_1^* = 0,$$

while

$$X_2^* = \begin{cases} U_2 & \text{if } U_1 \le \frac{1}{2}, \\ 1 + U_2 & \text{if } U_1 > \frac{1}{2}, \end{cases} \quad \text{but} \quad Y_2^* = 2\,U_2.$$

The second components are entirely different (X_2^* depends on U_1, while Y_2^* does not) and hence the Knothe–Rosenblatt distance does not *tend to* 0*,* $d_p^{KR}(\mu_n, \mu) \not\to 0$*, as* $n \to \infty$*.*

We conclude that 2 does *not* hold true and the topologies induced by the Knothe–Rosenblatt and the nested distances differ. More generally, we expect that there is no distinguished bicausal transport that can generate a topology compatible with the nested distance for $N > 1$.

7. Summary

In this article we investigated fundamental topological properties of the nested distance. In contrast to classical Wasserstein distances, for example, the nested distance cannot be characterized via integration on test functions, so that complete stochastic programs are the only distinguishing element of the topology induced by the nested distance. The nested distance is also not complete, again in contrast to the classical Wasserstein distance.

We obtained two main results. First, we demonstrated that the metric completion of the nested distance is the space of nested distributions with their classical Wasserstein metric, as introduced in [18]. This provides a connection between two hitherto unrelated mathematical objects. Second, we established that the topology generated by the nested distance is Polish, which we hope opens the way to future applications.

We finally introduced the Knothe–Rosenblatt distance between processes, which likewise takes into account the filtration structure. Its appeal lies in its apparent simplicity and the eminence of the Knothe–Rosenblatt rearrangement, as well as in the fact that in the one-dimensional setting this distance, and the nested distance, coincide. In this regard, we demonstrate that in higher dimensions, the topology generated by the Knothe–Rosenblatt distance is strictly stronger than the nested distance topology.

References

[1] B. Acciaio, J. Backhoff Veraguas, and A. Zalashko. Causal optimal transport and its links to enlargement of filtrations and continuous-time stochastic optimization. *ArXiv e-prints*, 2016.

[2] L. Ambrosio and N. Gigli. A user's guide to optimal transport. In *Modelling and optimisation of flows on networks*, volume 2062 of *Lecture Notes in Math.*, pages 1–155. Springer, Heidelberg, 2013.

[3] Luigi Ambrosio, Nicola Gigli, and Giuseppe Savaré. *Gradient Flows in Metric Spaces and in the Space of Probability Measures.* Birkhäuser Verlag, Basel, Switzerland, 2nd edition, 2005.

[4] J. Backhoff, M. Beiglböck, Y. Lin, and A. Zalashko. Causal transport in discrete time and applications. *preprint arXiv:1606.04062 [math.OC].*

[5] J. Backhoff Veraguas, M. Beiglböck, Y. Lin, and A. Zalashko. Causal transport in discrete time and applications. *ArXiv e-prints*, June 2016.

[6] Martin Barbie and Abhishek Gupta. The topology of information on the space of probability measures over Polish spaces. *J. Math. Econom.*, 52:98–111, 2014.

[7] M. Beiglböck, A. Cox, and M. Huesmann. Optimal transport and Skorokhod embedding. *Invent. Math., to appear*, 2016.

[8] M. Beiglböck, P. Henry-Labordère, and F. Penkner. Model-independent bounds for option prices—a mass transport approach. *Finance Stoch.*, 17(3):477–501, 2013.

[9] François Bolley. Separability and completeness for the Wasserstein distance. In Catherine Donati-Martin, Michel Émery, Alain Rouault, and Christophe Stricker, editors, *Séminaire de Probabilités XLI*, volume 1934 of *Lecture Notes in Mathematics*, pages 371–377. Springer, Berlin, Heidelberg, 2008.

[10] René Carmona, François Delarue, and Daniel Lacker. Mean field games with common noise. *Ann. Probab.*, 44(6):3740–3803, 2016.

[11] Y. Dolinsky and H. M. Soner. Martingale optimal transport and robust hedging in continuous time. *Probab. Theory Relat. Fields*, 160(1-2):391–427, 2014.

[12] A. Galichon, P. Henry-Labordère, and N. Touzi. A stochastic control approach to no-arbitrage bounds given marginals, with an application to lookback options. *Ann. Appl. Probab.*, 24(1):312–336, 2014.

[13] M. F. Hellwig. Sequential decisions under uncertainty and the maximum theorem. *J. Math. Econom.*, 25(4):443–464, 1996.

[14] Alexander S. Kechris. *Classical descriptive set theory*, volume 156 of *Graduate Texts in Mathematics.* Springer-Verlag, New York, 1995.

[15] R. Lassalle. Causal transference plans and their monge-kantorovich problems. arXiv:1303.6925, 2013.

[16] M. Nutz and F. Stebegg. Canonical Supermartingale Couplings. *ArXiv e-prints*, September 2016.

[17] G. Ch. Pflug and A. Pichler. Dynamic generation of scenario trees. *Comput. Optim. Appl.*, 62(3):641–668, 2015.

[18] Georg Ch. Pflug. Version-independence and nested distributions in multistage stochastic optimization. *SIAM Journal on Optimization*, 20:1406–1420, 2009.

[19] Georg Ch. Pflug and Alois Pichler. A distance for multistage stochastic optimization models. *SIAM Journal on Optimization*, 22(1):1–23, 2012.

[20] Georg Ch. Pflug and Alois Pichler. *Multistage Stochastic Optimization*. Springer Series in Operations Research and Financial Engineering. Springer, 2014.

[21] Georg Ch. Pflug and Alois Pichler. From empirical observations to tree models for stochastic optimization: Convergence properties. *SIAM Journal on Optimization*, 26(3):1715–1740, 2016.

[22] S.T. Rachev and L. Rüschendorf. *Mass transportation problems. Vol. I*. Probability and its Applications (New York). Springer-Verlag, New York, 1998.

[23] S.T. Rachev and L. Rüschendorf. *Mass transportation problems. Vol. II*. Probability and its Applications (New York). Springer-Verlag, New York, 1998. Applications.

[24] C. Villani. *Topics in optimal transportation*, volume 58 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI, 2003.

[25] C. Villani. *Optimal Transport. Old and New*, volume 338 of *Grundlehren der mathematischen Wissenschaften*. Springer, 2009.